COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
AT YALE UNIVERSITY

Box 2125, Yale Station
New Haven, Connecticut

COWLES FOUNDATION DISCUSSION PAPER NO. 205

EFFICIENT ESTIMATION WITH A PRIORI INFORMATION: A CLASSICAL APPROACH

Thomas J. Rothenberg

April 14, 1966

EFFICIENT ESTIMATION WITH A PRIORI INFORMATION: A CLASSICAL APPROACH

by

Thomas J. Rothenberg*

1. INTRODUCTION

A wide class of problems in econometric theory concern statistical inference in linear models when there are a priori constraints on the parameters. The famous "simultaneous equations problem" with identifying restrictions is perhaps the most notable example. Although these problems have received considerable attention over the past twenty years, there remain a number of gaps in the theory. In the present paper an attempt is made to develop some of the basic results of a general classical theory of estimation in the presence of a priori information and to apply these results to linear econometric models. In a subsequent paper the problem of efficiently estimating parameters of simultaneous equation systems will be examined as a special case of this general theory. There the value of overidentifying structural restrictions in increasing the efficiency of reduced form estimation will be analyzed in some detail.

The approach taken here is in the classical statistical tradition of evaluating estimating procedures on the basis of their sampling distributions

---

* The author is Assistant Professor of Economics at Northwestern University. This paper, a portion of a doctoral dissertation submitted to the Massachusetts Institute of Technology, was written in part while the author was a member of the Cowles Foundation research staff in Spring, 1965.

(or at least on the basis of approximations to these distributions).[1] Specifically, a "best" estimator is defined to be one which has smallest variance out of the class of all unbiased estimators using the available a priori information. Although this classical, minimum-variance-unbiased approach has been used in most previous studies in econometric theory, it is not completely satisfactory. Given the types of information economists possess and the uses they wish to make of their estimated models, it is likely that the classical statistical criteria are often inappropriate for econometrics. For many problems involving a priori information, a Bayesian decision-theoretic approach seems much more natural than the classical approach. Nevertheless, it is of considerable interest to see how a priori information can be incorporated into the classical theory, if for no other reason than to compare the classical results with the Bayesian results. The present paper thus ignores decision theoretic considerations and is completely "classical."

The concern of this paper is solely with the problem of optimal point estimation. The theory of optimal methods of testing hypotheses or of forming confidence regions is not discussed explicitly, although optimal estimates often form the basis for optimal tests and optimal confidence regions. There are two reasons for concentrating on point estimation. First, the theory of estimation can be presented at a very general, yet useful, level without getting involved in the details of special cases or in the difficult distribution

---

[1] The approach is classical in the sense that it is the one taught in the majority of postwar textbooks. Only a few decades ago, however, the word classical was used to describe the Bayesian view. Our definition seems to be consistent with modern usage.

theoretic problems which arise in the theory of hypothesis testing. Second, the most common decision-theoretic approaches to econometric problems are based on uses (e.g., forecasting) for which optimal decisions are essentially equivalent to optimal estimates. Hence the parallel between the classical and Bayesian approaches is made clearest via the theory of point estimation.

The literature on estimation under a priori constraints is rather limited. The only case that has been examined in great detail is that of the linear model under linear constraints.[1] The more general problem has been analyzed in a relatively few articles, most notably the early Cowles Commission studies [3, 20], a series of articles by Aitchison and Silvey [1,2, 24], and a section of the recent book by Malinvaud [22, Chapter 9]. Some of the results which follow are contained in the above-mentioned works, although their implications and generality seem not to have been previously explored.[2]

2. A PRIORI INFORMATION

One of the major problems in statistical inference is to develop methods of using sample information to obtain estimates of unknown parameters. However, there are many examples in the various areas of statistical application where the statistician possesses, in addition to the sample, a priori information about the parameters. For example, the econometrician may know

---

[1] See, for example, Theil [26, pp.331-33], Chipman and Rao [8], or Goldberger [13, pp. 255-65].

[2] The approach taken in the present paper is probably most similar to that of Hammersley [15] although his topic is quite different. The present paper is also closely related to a recent article by Klein [18].

from theoretical arguments that the marginal propensity to consume lies between zero and one or that a demand function is homogeneous of degree zero in prices and income. Our purpose is to explore the gain in efficiency which results from making use of such a priori information when estimating the parameters from a sample.

Let the vector of unknown parameters be denoted by θ. A priori information on the vector θ may be expressed in a number of different ways. It is useful to distinguish between two general classes of information -- stochastic information and nonstochastic information. The examples above concerning the aggregate consumption function and the demand function are illustrations of nonstochastic information since no random or probabilistic concepts are introduced. In the one case the set of possible values of θ (the marginal propensity to consume) is restricted to a finite interval rather than being allowed to assume any value on the real line. In the other case, the space of possible values for the various price and income coefficients is restricted to lie in a subspace defined by the homogeneity postulate. In general, nonstochastic information takes the form of restricting (making smaller) the set of possible values the vector θ may take. There are, of course, many ways to describe (or to generate) such a restricted set. Three simple methods which have particular relevance to econometrics are analyzed in the following sections.

Stochastic information is expressed by means of a random variable whose probability distribution involves θ. Such information may arise in either of two quite different ways. On the one hand, the statistician may

treat θ itself as though it were a random variable distributed according to some known subjective probability law. That is, the statistician expresses his "betting distribution," the odds he would give (or take) on gambles concerning the value of θ. On the other hand, the statistician may treat θ as a nonstochastic parameter but may possess an estimate $\hat{\theta}$ as a result of a previous sample. In such a case the probability distribution of $\hat{\theta}$ will generally depend on θ. In both cases the statistician must incorporate stochastic prior information with the new sample information. The former approach, of course, is Bayesian in spirit and does not fit into the classical framework of minimum-variance-unbiased estimation. The latter approach, however, is perfectly classical since it simply involves the combining of two samples (using a relative frequency interpretation of probability). Since the analysis in this paper is based on the classical theory, only the second type of stochastic information will be discussed.

The outline of the rest of the study is as follows. In the next section the classical theory of estimation without a priori information is summarized. In Sections 4-7, four different ways of expressing a priori information are presented and their impact on estimation efficiency examined. In a later paper some of these results will be applied to the case of estimating a system of simultaneous linear equations when there are nonstochastic overidentifying restrictions on the structure.

3. UNCONSTRAINED ESTIMATION

Since the purpose of this paper is to generalize some of the important classical theorems on efficient estimation so that they apply

to the case where a priori information concerning the parameters is available, it will be useful to begin by stating the classical theorems of unconstrained estimation. First it will be necessary to develop the required notation and assumptions. For ease of presentation, the analysis will be conducted under the assumption of independent sampling from identical probability distributions of the "continuous type." The generalizations to more general sampling schemes and to more general classes of probability distributions do not cause any particular difficulties. Let X_n be a vector random variable representing the sample outcome of n independent repetitions of an experiment. The joint probability function for X_n is assumed to be represented by a continuous density function

$$f_n(x, \theta) \tag{3.1}$$

where x represents the vector of observations and where θ is a vector of m unknown parameters.

The density function (3.1) is assumed to satisfy a number of regularity conditions. Since these conditions are discussed in detail elsewhere [9, pp. 500-501], they are merely summarized briefly here:

a) The set of possible values for θ, denoted by A, is an open subset of m-dimensional Euclidean space.

b) S, the set of x-values for which (3.1) is strictly positive, does not depend on θ.

c) The equation

$$\int f_n(x, \theta)dx = 1 \tag{3.2}$$

is satisfied for all θ in A . The integral in (3.2) is a multivariate integral over the finite-dimensional sample space S .

d) For all θ in A and almost every x , the functions $f_n(x, \theta)$ and $\log f_n(x, \theta)$ possess partial derivatives with respect to θ up to the third order. These derivatives are bounded by a function $K_1(x)$ which is finitely integrable over S . Furthermore, the third partial derivative of $\log f_n$ is bounded by a function $K_2(x)$ that has finite expected value. Hence, twice differentiation under the integral sign of (3.2) is possible.

e) The information matrix (defined below) is positive definite.

The above regularity conditions are needed in order to derive the classical theorems of estimation. Since these theorems are proven in the literature, most of the above conditions will not be explicitly used in the sequel. It may be noted in passing that most of the familiar distributions satisfy the above assumptions. (Discrete distributions can be included if the integrals are replaced by sums.) The most notable exceptions are the rectangular distribution and truncated distributions with the point of truncation depending on θ . Although it is possible to prove some of the classical theorems using weaker regularity conditions, no attempt to do so will be made here.

The traditional theory of estimation as developed by Fisher, Cramer, Rao and others, is concerned with finding efficient, or at least asymptotically efficient, estimators of the unknown parameter vector θ on the basis of the sample X_n . An estimator $t(X_n)$ is a vector of functions

which does not depend on the unknown θ . An estimator is unbiased if

$$E[t] \equiv \int t(x) f_n(x, \theta) dx = \theta \tag{3.3}$$

for every θ in A . An unbiased estimator t is efficient if its covariance matrix

$$V_t = [v_{ij}] = [E(t_i - \theta_i)(t_j - \theta_j)] \tag{3.4}$$

is at least as small as that of any other unbiased estimator. That is, t is efficient if, for every θ in A, $V_t - V_s$ is negative semidefinite for all unbiased estimators s .

Since it is often impossible to find an efficient estimator, it is useful to have an approximate concept. Consider a sequence of samples X_1 , X_2, ..., X_n (where the index refers to the sample size) and the corresponding sequence of density functions f_1, f_2, ..., f_n , where each f_i is a function of the same parameter vector θ . A sequence of estimators $t_1(X_1)$, $t_2(X_2)$, ..., $t_n(X_n)$ (where t_n represents the estimator based on a sample of size n) is <u>consistent</u> if,

$$\text{Plim } [t_n(X_n)] = \theta \ . \tag{3.5}$$

A consistent estimator t_n (or more precisely, a consistent estimator sequence $\{t_n\}$) is said to have an asymptotic covariance matrix V_t if, for all θ in A , the sequence of random variables $\{\sqrt{n}(t_n - \theta)\}$ converges in distribution to a random variable with mean zero and covariance matrix V_t . A consistent estimator t_n is said to be asymptotically efficient if $V_t - V_s$ is, for all θ in A , negative semidefinite for all consistent estimators s_n .

Asymptotic efficiency is usually defined only for a given class of consistent estimators (e.g. those satisfying certain regularity assumptions) since the class of all consistent estimators is not sufficiently well behaved. Note also that, whereas efficiency (for fixed n) is defined only for estimators with finite second moments, asymptotic efficiency requires only that the limiting distributions have finite second moments.

The basic theories of efficient estimation are usually stated in terms of the so-called information matrix. If X_n is a sample of size n and $f_n(x, \theta)$ is its density function, the information matrix is defined as

$$R_n = -E\left[\frac{\partial^2 \log f_n(X, \theta)}{\partial \theta_i \partial \theta_j}\right] = E\left[\frac{\partial \log f_n}{\partial \theta_i} \cdot \frac{\partial \log f_n}{\partial \theta_j}\right] \tag{3.6}$$

where the last equality is verified by differentiating equation (3.2). The asymptotic information matrix associated with a sequence of samples $X_1, X_2, \ldots$, and a sequence of densities $f_1, f_2, \ldots$, is defined as

$$R = \lim_{n\to\infty} \frac{1}{n} R_n \,. \tag{3.7}$$

It will be assumed that for every θ in A, R and each R_n exist and are positive definite.

Two major results of classical estimation theory may now be stated:

THEOREM A. The matrix R_n^{-1} is a lower bound for the covariance matrix of any unbiased estimator of θ. There exists an unbiased estimator whose variance attains this bound if and only if the logarithmic derivative of the likelihood function [1] takes the form

---

[1] We shall use the phrases "likelihood function" and "joint density function" interchangeably when referring to $f_n(x, \theta)$.

$$(3.8) \qquad \frac{\partial \log f_n}{\partial \theta} = R_n(t - \theta)$$

where t, a vector not depending on θ, is the minimum variance bound (MVB) estimator.

THEOREM B. The matrix R^{-1} is "essentially" a lower bound for the asymptotic covariance matrix of any consistent estimator of θ. Furthermore, this lower bound is attained by the maximum likelihood estimator.

These two results have a long history and are associated with such statisticians as H. Cramer, D. Dugue, R. A. Fisher, M. Frechet, and C. R. Rao. For convenience, Theorem A will sometimes be referred to as the Cramer-Rao inequality and Theorem B as the asymptotic Cramer-Rao inequality.[1]

The remainder of this paper will deal with generalizations and applications of these two inequalities when a priori information is available. First, however, some comments concerning these classical theorems are in order. It is clear that equation (3.8) is quite restrictive and that only for a few density functions will the bound R_n^{-1} be attainable.[2] Moreover, if the bound is attainable for one set of parameters θ, it will in general not be attainable for any new set of parameters obtained by nonlinear transformation. (For example, in the one-parameter case, if the lower bound is attainable for θ, the bound is not attainable for θ^2.) Since the bound R_n^{-1} is not usually a best lower bound, it may reasonably be asked whether the bound is worth much attention. The answer to the question lies in the fact that, for large n,

---

[1] For proofs of the two classical theorems see Kendall and Stuart [16, pp. 8-60] and the references cited there.

[2] The bound will be attainable only if there exists a set of n sufficient statistics. This will be the case only if f_n is a member of the Pittman-Koopman class of densities.

the difference between the best possible bound and R_n^{-1} is very small. Hence the Cramer-Rao inequality gives an approximate expression for the lowest attainable variance of an unbiased estimator. The assumption of unbiasedness is crucial to the small-sample classical theory since uniformly best estimators without the unbiasedness constraint do not exist. For large n, however, this constraint is unimportant since the distribution of an asymptotically efficient estimator can always be approximated by a distribution with mean θ.

The asymptotic Cramer-Rao inequality is a statement of the above-mentioned approximation. Although R_n^{-1} may not be attainable for finite n, R^{-1} is always attainable in infinite samples. The word "essentially" in the statement of the second theorem allows for certain pathological cases where covariance matrices, for a few values of θ, are smaller than R_n^{-1}. These cases can be eliminated by stating regularity conditions on the class of estimators considered or by redefining efficiency by changing the phrase "for all θ" to the phrase "for all θ except for a set of measure zero." A precise statement of the asymptotic inequality where careful account is taken of the conditions needed for its validity can be found in LeCam [21]. Our concern is the generalization of the classical theorems for non-pathological cases and hence all of the results that follow should be understood to require the same qualifications as discussed here. We begin by examining the increase in efficiency which is possible when stochastic information from a previous sample is present.

4. CASE I: PREVIOUS SAMPLE

4.1. The Minimum Variance Bound

Suppose that the statistician has available to him, in addition to the sample X_n, another independent sample from which he obtains the estimator $\hat{\theta}$. Suppose further that the statistician knows that $\hat{\theta}$ is distributed according to the probability density $f_o(\hat{\theta}, \theta)$. That is, the probability law for $\hat{\theta}$ is a known function of the unknown parameter θ. Then the joint density function for X_n and $\hat{\theta}$ is given by

$$f(\hat{\theta}, x; \theta) = f_n(x; \theta) \cdot f_o(\hat{\theta}, \theta) \tag{4.1}$$

as long as X_n and $\hat{\theta}$ are independently distributed.

If f_o and f_n satisfy all the regularity conditions of Section 3 then f will also. Thus, all the assumptions of the Cramer-Rao inequality are satisfied if one treats f as the density function for the sample $(X_n, \hat{\theta})$. If the information matrix for f is denoted by $\bar{R}_n$, it follows from the multiplicative form of (4.1) that

$$\bar{R}_n = R_n + R_o \tag{4.2}$$

where R_o is the information matrix associated with f_o. Since all three matrices are positive definite, it follows that $R_n^{-1} - \bar{R}_n^{-1}$ is positive definite.[1] Hence we have the result that the lower bound for the variance

---

[1] See Appendix A.

of an unbiased estimator is decreased when information from a previous sample is used.

The question of the attainability of the bound in finite samples is difficult to answer in general. Although it does not appear possible to specify easily interpretable necessary conditions for the attainability of the lower bound, an interesting set of sufficient conditions can be established. Recall that the Cramer-Rao bound can be attained if and only if f can be expressed in the form

$$\frac{\partial \log f}{\partial \theta} = \bar{R}_n(t^* - \theta) \tag{4.3}$$

where t^* is an estimator independent of θ. Suppose that f_n and f_o can be written as

$$\begin{aligned} \frac{\partial \log f_n}{\partial \theta} &= R_n(t - \theta) \\ \frac{\partial \log f_o}{\partial \theta} &= R_o(s - \theta) \end{aligned} \tag{4.4}$$

where t depends on X_n and s depends on $\hat{\theta}$. That is, suppose that t is a MVB estimator of θ when only X_n is available and that s is a MVB estimator when only $\hat{\theta}$ is available.

An expression for the logarithmic derivative of f is obtained by adding the two equations of (4.4):

$$\begin{aligned} \frac{\partial \log f}{\partial \theta} &= \frac{\partial \log f_n}{\partial \theta} + \frac{\partial \log f_o}{\partial \theta} \\ &= (R_n + R_o)[(R_n + R_o)^{-1}(R_n t + R_o s) - \theta] \\ &= \bar{R}_n(t^* - \theta) \, . \end{aligned} \tag{4.5}$$

Thus we have the required form as long as

$$t^* = (R_n + R_o)^{-1}(R_n t + R_o s) \tag{4.6}$$

does not depend on θ. It is clear that t^* will be independent of θ if R_n and R_o do not depend on θ. This is overly strong, however; it is sufficient that R_n and R_o each factors into a matrix independent of θ and a common scalar which may depend on θ.

The above discussion may be summarized in the following extension of Theorem A:

THEOREM 1. In the presence of stochastic prior information expressed by an independent estimator of θ, the matrix $\bar{R}_n^{-1}$ defined by (4.2) is a lower bound for the covariance matrix of any unbiased estimator of θ. The bound, however, is attainable only under restrictive conditions. One set of sufficient conditions is that both f_n and f_o can be written in the form (3.8) with R_n and R_o not depending on θ (except perhaps for identical scalar multiples).

4.2. An Example

Consider the normal regression model

$$y = X\beta + u \tag{4.7}$$

where y is an n-dimensional vector of observations on a random variable, X is a $n \times m$ matrix of nonstochastic variables, β is a vector of m unknown parameters and u is a vector of n independent normal random

errors with zero mean and covariance matrix $\sigma^2 I$. The likelihood function for the sample is

$$(4.8) \qquad f_n(y, \beta, \sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}n} \exp\{-\tfrac{1}{2}\sigma^{-2}(y - X\beta)'(y - X\beta)\} .$$

The logarithmic derivatives are

$$(4.9) \qquad \frac{\partial \log f_n}{\partial \beta} = \frac{1}{\sigma^2} X'X\,(b - \beta)$$

$$(4.10) \qquad \frac{\partial \log f_n}{\partial \sigma^2} = \frac{n}{2\sigma^4}\left(\frac{u'u}{n} - \sigma^2\right)$$

where $b = (X'X)^{-1}X'y$ is the least-squares estimator. The information matrix for β and σ^2 is the $(m + 1) \times (m + 1)$ matrix

$$(4.11) \qquad R_n = \begin{bmatrix} \frac{1}{\sigma^2} X'X & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}$$

Because R_n is block diagonal, the Cramer-Rao bound for β may be examined separately from that for σ^2 . It is apparent that (4.9) is of the form (3.8); however, equation (4.10) is not since $u'u/n$ depends on the parameter β . Thus, as is usual with the normal distribution, the Cramer-Rao bound is not attainable for σ^2 . No unbiased estimator has a variance as low as $2\sigma^4/n$. However, the least-squares estimator b does have the covariance matrix $\sigma^2(X'X)^{-1}$ and hence is a MVB estimator of β .

Suppose now that there exists a previous sample <u>from the same process</u>. That is, the statistician has available an observation on an n_o-dimensional normal random vector y_o and an $n_o \times m$ matrix X_o such that $E[y_o] = X_o\beta$ and $Var[y_o] = \sigma^2 I$. The likelihood function for the sample y_o will be of the same form as (4.8) with subscripts on y, n, and X. The joint sample (y, y_o) will have a normal density f such that

$$(4.12) \qquad \frac{\partial \log f}{\partial \beta} = -\frac{1}{\sigma^2} X'X(b - \beta) + \frac{1}{\sigma^2} X_o'X_o(b_o - \beta)$$

$$= \frac{1}{\sigma^2}(X'X + X_o'X_o)(b^* - \beta)$$

$$(4.13) \qquad \frac{\partial \log f}{\partial \sigma^2} = \frac{n + n_o}{2\sigma^4}\left[\frac{u'u + u_o'u_o}{n + n_o} - \sigma^2\right]$$

where $b^* = (X'X + X_o'X_o)^{-1}(X'y + X_o'y_o)$. Hence b^* is an MVB estimator of β with covariance matrix equal to the bound $\sigma^2(X'X + X_o'X_o)^{-1}$, the northwest submatrix of

$$(4.14) \qquad \bar{R}_n^{-1} = \begin{bmatrix} \sigma^2 (X'X + X_o'X_o)^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n + n_o} \end{bmatrix} .$$

The combined sample gives rise to an efficient estimator for β and an attainable Cramer-Rao bound because both samples come from the same stochastic process. If, however, the second sample came from a process with a different variance, b^* would depend on the unknown σ^2 and the

bound would not be attainable. Prior information which is derived from a different stochastic process than that which produces the sample will in general not give rise to an estimator whose variance equals the lower bound.

4.3. The Asymptotic Bound

In order to complete the analysis of stochastic prior information, our attention must turn to the asymptotic case and an extension of Theorem B. Defining the asymptotic information matrix for f by

$$\bar{R} = \lim 1/n\ \bar{R}_n = \lim 1/n\ [R_n + R_o]\ , \tag{4.15}$$

we can argue as above that, since $\bar{R}$ satisfies the same regularity conditions as R, the classical theorem applies. The only remaining question is the value of $\bar{R}$. This, of course, depends on the value of

$$\lim 1/n\ R_o \tag{4.16}$$

If it is supposed that the prior information does not change as the sample X_n gets larger, then R_o is a fixed matrix and the limit in (4.16) is simply the zero matrix. As the sample gets larger, the prior information plays a smaller role. In the limit, it is of absolutely no value.

A more interesting case is to assume that the prior information is of the same order of magnitude as the sample information. In other words, it can be assumed that the limit in (4.16) is a positive definite matrix $\bar{R}_o$. This assumption should be interpreted as follows: We are interested in an approximation to $\bar{R}_n$ which is valid for "large" n. By "large" one means a

sample size small enough to occur in practice, but large enough to make the approximation error reasonably small. For such a sample, R_o/n may very well be much larger than the approximation error. In such cases it is convenient to accept the fiction that R_o is a function of n and that (4.16) possesses a nonzero limit $\bar{R}_o$. Then one can conclude that stochastic information reduces the bound for the asymptotic covariance matrix of a consistent estimator. These results can be summarized in the following:

THEOREM 2. In the presence of stochastic prior information expressed by an independent estimator, the matrix $\bar{R}^{-1}$ defined in (4.15) is essentially a lower bound for the asymptotic covariance matrix of any consistent estimator of θ. This lower bound is attained by the maximum-likelihood estimator (where f is the relevant likelihood function). The matrix $\bar{R}^{-1}$ differs from R^{-1} only if the prior information is of the same order of magnitude as the sample information. If the prior information is independent of n there is no gain in asymptotic efficiency.

5. CASE II: CONSTRAINT EQUATIONS

5.1 The Minimum Variance Bound

One of the easiest ways of expressing nonstochastic prior information is by means of a set of equations which θ is constrained to satisfy. For example, if θ is the vector of all price and income elasticities of demand in a many-commodity market, the theory of utility maximization subject to constraint implies that certain weighted sums of these elasticities must equal zero. In general, suppose that the set of possible values that the unknown

parameter vector θ may take is restricted to A_g, the solution set of k equations

$$(5.1) \qquad g_i(\theta) = 0 \qquad (i = 1, \ldots, k)$$

where k is less than m, the number of unknown parameters. Suppose further that the g_i are continuous and possess partial derivatives of at least the second order. It will be assumed that the matrix of first partial derivatives

$$(5.2) \qquad G = [g_{ij}] = \left[\frac{\partial g_i}{\partial \theta_j}\right]$$

has full row rank k when evaluated at the true parameter θ^o. That is, the equations (5.1) are functionally independent in a neighborhood of θ^o.

The derivation of a lower bound for the variance of an unbiased estimator proceeds as follows. Since A_g is not an open set in m-space, Theorem A is not applicable. However, it can easily be modified. For any unbiased estimator t

$$(5.3) \qquad \int (t_i - \theta_i) f_n(x, \theta)\, dx = 0 \qquad (i = 1, \ldots, m)$$

for all θ in A_g, the solution set of (5.1). Let θ^o be the true parameter and θ^1 be another vector in A_g. Then, for all i,

$$(5.4) \qquad \int (t_i - \theta_i^o)[f_n(\theta^1, x) - f_n(\theta^o, x)]dx = \theta_i^1 - \theta_i^o .$$

Defining $y_i = \theta_i^1 - \theta_i^o$ and using the mean value theorem, we can write

$$(5.5) \qquad \int (t_i - \theta_i^o)\Big[\sum_j \frac{\partial f_n}{\partial \theta_j} y_j\Big]dx = y_i$$

where the partial derivatives are evaluated at θ^*, a point between θ^1 and θ^0. Then, turning to logarithms and defining $c = (c_1, \ldots, c_m)'$ to be an arbitrary nonstochastic vector, we have

$$\int \sum_i c_i(t_i - \theta_i^0) \sqrt{f_n} \cdot \sum_j \frac{\partial \log f_n}{\partial \theta_j} y_j \sqrt{f_n}\, dx = \sum_i c_i y_i \tag{5.6}$$

where f_n is also evaluated at θ^*. Application of the Cauchy-Schwartz inequality to (5.6) yields

$$c'V^*c \cdot y'R_n^*y \geq (c'y)^2 \tag{5.7}$$

where V^* is the second-moment matrix of t around θ^0 using the density $f_n(x, \theta^*)$ and R_n^* is the information matrix evaluated at θ^*. The inequality (5.7) must hold for all y such that θ^1 is in A_g. But, application of the mean value theorem to (5.1) yields

$$G^{**}y = 0 \tag{5.8}$$

where G^{**} is the Jacobian matrix G evaluated at some vector θ^{**} between θ^0 and θ^1. As $|y|$ approaches zero, V^* approaches the true covariance matrix of t, R_n^* approaches the true information matrix R_n, and G^{**} approaches G^0. Since (5.7) must hold for all θ^1 is A_g, it must also hold in the limit as $|y|$ goes to zero (along a path in A_g). Hence, after dividing the inequality (5.7) and the equation (5.8) by $|y|$, one finds that in the limit the following inequality holds:

$$c'Vc \geq \frac{(c'y)^2}{y'R_n y} \tag{5.9}$$

for all y satisfying the constraint

$$G^{o}y = 0 \ . \tag{5.10}$$

A lower bound for $c'Vc$ is obtained by finding the maximum value the right hand side of (5.9) may take. This leads to the following extremal problem:

$$\begin{aligned} &\underset{y}{\text{maximize}} \quad (c'y)^2 \\ &\text{subject to} \\ &G^{o}y = 0 \\ &y'R_n y = 1 \ . \end{aligned}$$

This problem is easily solved using the method of Lagrange multipliers. The inequality (5.9) becomes $c'Vc \geq c'P_n c$, where P_n is given by[1]

$$P_n = R_n^{-1} - R_n^{-1}G'(GR_n^{-1}G')^{-1}GR_n^{-1} \ , \tag{5.11}$$

a matrix having rank $m - k$.[2] This maximum is attained when y is any multiple of $P_n c$.

This result can be expressed more simply in terms of the bordered information matrix

$$\begin{bmatrix} R_n & G' \\ G & 0 \end{bmatrix} \tag{5.12}$$

---

[1] For notational convenience the superscript on G^{o} is dropped. In the sequel all derivatives are evaluated at the true parameter θ^{o} unless otherwise stated.

[2] See Appendix A.

and its conformably partitioned inverse

$$\begin{bmatrix} R_n & G' \\ G & 0 \end{bmatrix}^{-1} = \begin{bmatrix} P_n & * \\ * & * \end{bmatrix} . \tag{5.13}$$

Thus the Cramer-Rao bound for the variance of an unbiased estimator when prior constraints are used is given by the $m \times m$ northwest submatrix of the inverted bordered information matrix. The decrease in the bound due to the prior information is given by

$$R_n^{-1} - P_n = R_n^{-1}G'(GR_n^{-1}G')^{-1}GR_n^{-1} , \tag{5.14}$$

a positive semidefinite matrix having rank equal to the number of constraints k.[1]

Given the constraint equations, it is possible to relax the assumption that R_n is nonsingular. The extremal problem just solved is the same as the problem

$$\underset{y}{\text{maximize}} \quad (c'y)^2$$

subject to

$$Gy = 0$$

$$y'(h_n + G'G)y = 1 .$$

If the matrix $R_n + G'G$ is nonsingular, then equations (5.11)-(5.14) remain valid if R_n is replaced by $R_n + G'G$. The lower bound is then given by the appropriate submatrix of

---

[1] See Appendix A.

$$(5.15) \qquad \begin{bmatrix} R_n + G'G & G' \\ G & 0 \end{bmatrix}^{-1} = \begin{bmatrix} \overline{P}_n & * \\ * & * \end{bmatrix} .$$

In this case the availability of a priori information makes estimation possible. Without the information the parameters could not be estimated at all.

The bound P_n is, of course, attainable only under restrictive conditions. Typically, no unbiased estimator can be found with a covariance matrix equal to P_n. From the above derivation, it is seen that the bound is attainable only if the Cauchy-Schwartz inequality when applied to (5.6) remains an equality for $y = P_n c$ and all c. This will occur if and only if, for all x, f_n is of the form

$$(5.16) \qquad P_n \frac{\partial \log f_n}{\partial \theta} = t - \theta$$

where t does not depend on θ. Although it does not seem possible to state more useful necessary conditions for this to hold, it is possible to present an interesting set of sufficient conditions.

Suppose that there exists a MVB estimator when there is no a priori information. This means that the likelihood function can be written in the form (3.8). Suppose further that the constraints (5.1) are linear so that they take the form $G\theta = a$. In that case one can write

$$P_n \frac{\partial \log f_n}{\partial \theta} = [I - R_n^{-1}G'(GR_n^{-1}G')^{-1}G]R_n^{-1} \frac{\partial \log f_n}{\partial \theta}$$

$$(5.17) \qquad = [I - R_n^{-1}G'(GR_n^{-1}G')^{-1}G](s - \theta)$$

$$= [s - R_n^{-1}G'(GR_n^{-1}G')^{-1}(Gs - a) - \theta]$$

where neither G nor s depends on θ. Then, if R_n is independent of θ except for at most a scalar multiple, (5.17) is of the form needed for an attainable bound. The MVB estimator is given by

$$s^* = s - R_n^{-1}G'(GR_n^{-1}G')^{-1}(Gs - a) . \tag{5.18}$$

The results obtained above may be summarized as follows:

THEOREM 3. In the presence of prior knowledge expressed by a set of k independent constraint equations, the matrix P_n given in (5.11) is a lower bound for the covariance matrix of any unbiased estimator of θ. The possible efficiency gain $R_n^{-1} - P_n$ is a positive semidefinite matrix of rank k. Sufficient (but not necessary) conditions for P_n to be attainable are

a) R_n^{-1} is attainable when the constraints are ignored,

b) R_n can be written as the product of a matrix which does not depend on θ and a scalar which may depend on θ,

c) The constraints $g_i(\theta)$ are linear.

5.2. The Asymptotic Bound

The asymptotic extension of Theorem 3 is straightforward. It is merely necessary to define

$$P = \lim_{n\to\infty} \frac{1}{n} P_n = R^{-1} - R^{-1}G'(GR^{-1}G')^{-1}GR^{-1} \tag{5.19}$$

which is the $m \times m$ northwest submatrix of

$$\lim_{n\to\infty} \begin{bmatrix} \frac{1}{n} R_n & G' \\ G & 0 \end{bmatrix}^{-1} = \begin{bmatrix} R & G' \\ G & 0 \end{bmatrix}^{-1}. \tag{5.20}$$

Then one can state:

THEOREM 4. In the presence of prior information expressed by a set of constraint equations, the matrix P is essentially a lower bound for the asymptotic covariance matrix of any consistent estimator of θ. The efficiency gain $R^{-1} - P$ is a positive semidefinite matrix having rank k. The bound is attained by the constrained maximum-likelihood estimator.

A complete proof of Theorem 4 is very lengthy and difficult. Fortunately, however, the classical proofs of Theorem B can be applied with only minor modification. It is necessary to show that the likelihood function defined on the restricted parameter space A_g satisfies the regularity assumptions assumed by LeCam [21]. This has been done by Aitchison and Silvey [2] in the course of deriving the asymptotic distribution of the constrained maximum-likelihood estimator.

The method of maximum likelihood is by no means unique in giving estimators with optimal large-sample properties. Another general principle of estimation -- the method of minimum chi-square -- also gives rise to asymptotically efficient estimators.[1] Consider the quadratic form

---

[1] The minimum-chi-square method or some variant of it is used by Malinvaud [22, pp. 283-86], who refers to it as the minimum distance method, and by Basmann [4,5] who refers to it as the generalized classical estimating method. For further discussion of the general principle see Ferguson [10].

$$\varphi(\theta) = (t - \theta)'\hat{R}(t - \theta) \tag{5.21}$$

where t is an estimator of θ which is asymptotically normal and efficient when there are no constraints; $\hat{R}$ is the asymptotic information matrix R evaluated at $\theta = t$. The estimator t might be, for example, the unconstrained maximum-likelihood estimator. In any case, t is an estimator which is asymptotically normal with mean θ and covariance matrix R^{-1}. The variable $n\varphi$ converges in distribution to a chi-square variate as n approaches infinity. The minimum-chi-square estimator of θ in the presence of the prior information is given by $\hat{\theta}$, the solution to the following extremal problem:

$$\min_{\theta} (t - \theta)'\hat{R}(t - \theta)$$

(5.22)

subject to

$$g(\theta) = 0 .$$

The linearized minimum chi-square estimator is given by $\hat{\theta}^*$, the solution to the modified extremal problem:

$$\min_{\theta} (t - \theta)'\hat{R}(t - \theta)$$

(5.23)

subject to

$$g(t) + \hat{G}(\theta - t) = 0$$

where $\hat{G}$ is the Jacobian matrix G evaluated at $\theta = t$. The solution to the first extremal problem cannot be given explicitly. The solution to the second problem, however, is easily obtained as

$$\hat{\theta}^* = t - R^{-1}G'(GR^{-1}G')^{-1}g(t) \tag{5.24}$$

where, for typographical reasons, we have omitted the "hat" from R and G.

The basic theorem of the minimum-chi-square method is that both $\hat{\theta}$ and $\hat{\theta}^*$ are asymptotically efficient. We shall indicate the proof for the case of $\hat{\theta}^*$, referring the reader to Chiang [7] and Ferguson [10] for the complete proof. Using the mean value theorem, we can write (5.24) as

$$\hat{\theta}^* - \theta^\circ = [I - R^{-1}G'(GR^{-1}G')^{-1}G^*](t - \theta^\circ) \tag{5.25}$$

where G^* is G evaluated at some point between t and θ°. Let the expression in square brackets be denoted by A. Then $\operatorname{Plim} A = PR$.

Hence $n^{\frac{1}{2}}(\hat{\theta}^* - \theta^\circ)$ has the same asymptotic distribution as $PR\, n^{\frac{1}{2}}(t - \theta^\circ)$. But the latter random variable is asymptotically normal with mean zero and covariance matrix

$$PRR^{-1}RP = P$$

Hence $\hat{\theta}^*$ has an asymptotic variance equal to the Cramer-Rao lower bound.

It is important to emphasize that the optimality of the minimum-chi-square estimator depends crucially on the assumption that the unconstrained estimator t has an asymptotic covariance matrix equal to R^{-1}. Only if such an estimator t, can be easily calculated will the minimum-chi-square approach

be a practical method. Fortunately, for many problems met in practice an easy-to-calculate, efficient estimator for the unconstrained problem exists.

5.3. An Example

An important example where the bound P_n is attainable in finite samples is the normal linear regression model with linear constraints.[1] Consider again the regression equation

$$y = X\beta + u \tag{5.26}$$

where y is an n-dimensional vector of observations on a random variable, X is a $n \times m$ matrix of nonstochastic variables, β is a vector of m unknown parameters, and u is a vector of n independent normal random errors with zero mean and constant variance σ^2.

Suppose that the prior constraints are of the form

$$G\beta = a . \tag{5.27}$$

Since the information matrix for (β, σ^2) is block diagonal and the constraints do not involve σ^2, attention can be focused solely on β. The information matrix for β is

$$R_n = \sigma^{-2}(X'X) \tag{5.28}$$

and the logarithmic derivative of the likelihood function is

$$\frac{\partial \log f_n}{\partial \beta} = \sigma^{-2}(X'X)(b - \beta) \tag{5.29}$$

---

[1] This case has been treated by Theil [26, pp. 331-33] and by Chipman and Rao [8].

where b is the least-squares estimator $(X'X)^{-1}X'y$. The constrained least-squares estimator is found by the use of Lagrange multipliers to be

$$\hat{\beta} = b - (X'X)^{-1}G'[G(X'X)^{-1}G']^{-1}(Gb - a) \tag{5.30}$$

and the covariance matrix for $\hat{\beta}$ is

$$P_n = \sigma^2\{(X'X)^{-1} - (X'X)^{-1}G'[G(X'X)^{-1}G']^{-1}G(X'X)^{-1}\} . \tag{5.31}$$

Hence $\hat{\beta}$ is a MVB estimator in the presence of the a priori information. Under plausible conditions on $X'X$, $\hat{\beta}$ is also asymptotically efficient; but that is a much weaker result. It is easy to verify that the constrained least-squares estimator $\hat{\beta}$ is also the constrained maximum-likelihood and minimum-chi-square estimator.

6. CASE III: CONSTRAINT PARAMETERS

6.1. The Minimum Variance Bound for θ

A third way of expressing prior information is to assume that the elements of θ are related functionally to another set of parameters. Let α be a vector of r unknown parameters. Suppose that the statistician knows that each θ_i is a given function of the elements of α. That is,

$$\theta_i = h_i(\alpha) \qquad (i = 1, \ldots, m) \tag{6.1}$$

where each h_i possesses bounded partial derivatives of at least the second order. It is assumed that the matrix of first partial derivatives

$$H = [h_{ij}] = \left[\frac{\partial h_i}{\partial \alpha_j}\right] \tag{6.2}$$

has constant rank ρ in an open neighborhood containing the true parameter α^o. It is further assumed that the set of possible values that α may take is an open set in r-dimensional Euclidean space.

These assumptions are strong enough to insure that, in a neighborhood of θ^o, $m - \rho$ of the θ_i can be expressed as functions of the remaining ρ. That is, it is possible to convert the m equations (6.1) and the r new variables into $m - \rho$ constraint equations involving only the θ_i. Hence, the case being considered here is essentially equivalent to the case presented in the previous section. However, for a number of problems (of which the simultaneous equations problem is an important example), the form (6.1) is more natural and more easily interpreted than the derived constraint equations. Furthermore, it will be possible here to derive covariance matrices for estimates of both α and θ.

We begin by assuming that H has full row rank r. The assumption will be dropped later. The derivation of the variance bound for this case can be split into two parts. First, a lower bound for V_θ, the covariance matrix of an unbiased estimator of θ, is obtained; then a lower bound for V_α, the covariance matrix of an unbiased estimator of α, is obtained. For the first part, much of the derivation of Section 5 is relevant. Equations (5.4) - (5.7) are still valid except that the values of y (that is, $\Delta\theta$) for which they hold now differ. Instead of the set A_g, θ^1 is now constrained to lie in A_h, the solution set of (6.1). The same limit argument applies as before since h is assumed to be continuously differentiable. The basic inequality (5.9) remains valid but (5.10) now must read

(6.3) $$y = H^{o}z$$

where z (representing $\Delta\alpha$) may take any value in the neighborhood of the origin in r-space. The lower bound for $c'V_{\theta}c$ is obtained by solving the constrained extremal problem:[1]

$$\max_{y} (c'y)^2$$

subject to

$$y = Hz$$

$$y'R_n y = 1$$

z unrestricted.

This, however, is the same as the problem:

$$\max_{z} (c'Hz)^2$$

subject to

$$z'H'R_n Hz = 1$$

The solution to the latter problem is obtained using the method of Lagrange multipliers. The objective function is maximized when $z = N_n c$ where

(6.4) $$N_n = H(H'R_n H)^{-1}H'$$

is a matrix having rank r. The value of the objective function at the maximum is $c'N_n c$. Hence, the Cramer-Rao inequality becomes

---

[1] Again, for notational convenience, we drop the superscript on H^{o}.

$$c'V_\theta c \geq c'H(H'R_nH)^{-1}H'c \quad . \tag{6.5}$$

The decrease in the bound due to the constraints, $R_n^{-1} - N_n$, is shown in Appendix A to be a positive semidefinite matrix of rank $m - r$.

6.2. The Minimum Variance Bound for α

The second part of the problem is to find a lower bound for V_α. Here the procedure is quite simple. Since the functions f_n and h are continuously differentiable, the compound function

$$f_n[x, h(\alpha)] = f_n^*(x, \alpha) \tag{6.6}$$

is a density function which satisfies the regularity assumptions of Section 3. Hence the information matrix for α is obtained by the chain rule:

$$\frac{\partial \log f_n^*}{\partial \alpha_i} = \sum_k \frac{\partial \log f_n}{\partial \theta_k} \cdot \frac{\partial \theta_k}{\partial \alpha_i} \tag{6.7}$$

$$\frac{\partial^2 \log f_n^*}{\partial \alpha_i \, \partial \alpha_j} = \sum_k \sum_p \frac{\partial^2 \log f_n}{\partial \theta_k \, \partial \theta_p} \frac{\partial \theta_k}{\partial \alpha_i} \frac{\partial \theta_p}{\partial \alpha_j} + \sum_k \frac{\partial \log f_n}{\partial \theta_k} \frac{\partial^2 \theta_k}{\partial \alpha_i \, \partial \alpha_j} \quad . \tag{6.8}$$

The second term on the right of (6.8) has zero expectation since

$$\int \frac{\partial \log f_n}{\partial \theta_k} f_n = \int \frac{\partial f_n}{\partial \theta_k} = \frac{\partial}{\partial \theta_k} \int f_n = \frac{\partial}{\partial \theta_k} 1 = 0 \quad .$$

Thus

$$(6.9) \qquad -E\,\frac{\partial^2 \log f_n^*}{\partial\alpha_i\,\partial\alpha_j} = \sum_k \sum_p h_{ki} r_{kp} h_{pj}$$

and, if H has rank r, the inverted information matrix for α is

$$(6.10) \qquad M_n = (H'R_nH)^{-1} .$$

The Cramer-Rao inequality becomes

$$(6.11) \qquad c'V_\alpha c \geq c'M_n c = c'\,(H'R_nH)^{-1}c .$$

There exists an unbiased estimator whose variance equals the lower bound only under certain restrictive conditions. Again, linearity of the constraints provides the simplest example. Suppose t is a MVB estimator when no constraints are present and hence

$$(6.12) \qquad \frac{\partial \log f_n}{\partial\theta} = R_n(t - \theta) .$$

If the constraints (6.1) are linear so that they are of the form

$$(6.13) \qquad \theta = H\alpha + a ,$$

one can write

$$(6.14) \qquad N_n \frac{\partial \log f_n}{\partial\theta} = N_n R_n(t - H\alpha - a)$$

$$= [H(H'R_nH)^{-1} H'R_n(t - a) + a] - \theta$$

and, using (6.7),

$$(6.15) \qquad M_n \frac{\partial \log f_n^*}{\partial \alpha} = M_n H' \frac{\partial \log f_n}{\partial \theta}$$

$$= (H'R_nH)^{-1}H'R_n(t - H\alpha - a)$$

$$= [(H'R_nH)^{-1}H'R_n(t - a)] - \alpha .$$

But, analogous to (5.16), the bound is attainable if the right hand sides of (6.14) and (6.15) are the difference between the estimator and the parameter. This is the case if R_n does not depend on α (except for perhaps a scalar multiple).

6.3 The Finite-sample Theorems

In summary we can state

THEOREM 5. In the presence of prior information expressed by a set of constraint parameters α which are related to θ by known differentiable equations (6.1), the matrix N_n given in (6.4) is a lower bound for the covariance matrix of any unbiased estimator of θ and the matrix M_n given in (6.10) is a lower bound for the covariance matrix of any unbiased estimator of α. The possible efficiency gain $R_n^{-1} - N_n$ is a positive semi-definite matrix of rank $m - r$. Sufficient (but not necessary) conditions for M_n and N_n to be attainable are

a) R_n^{-1} is attainable when the constraints are ignored,

b) R_n can be written as the product of a matrix which does not depend on θ and a scalar which may depend on θ,

c) the constraints $h_i(\alpha)$ are all linear.

Again these results can be expressed more simply in terms of a bordered information matrix. Consider the square matrix of order $2m + r$

$$(6.16) \qquad \mathcal{R}_n = \begin{bmatrix} R_n & 0 & -I \\ 0 & 0 & H' \\ -I & H & 0 \end{bmatrix}$$

which is partitioned into three row blocks (the first containing m rows, the second containing r rows, and the third containing m rows) and similarly three column blocks. This matrix turns out to be minus the expected value of the second partial derivative matrix of the Lagrangean

$$(6.17) \qquad L(\theta, \alpha, \lambda) = \log f_n(x, \theta) + \sum_{i=1}^{m} \lambda_i[\theta_i - h_i(\alpha)]$$

where the λ_i are taken to have zero expected value.[1] That is,

$$(6.18) \qquad \mathcal{R}_n = -E \begin{bmatrix} \frac{\partial^2 L}{\partial\theta\partial\theta'} & \frac{\partial^2 L}{\partial\theta\partial\alpha'} & \frac{\partial^2 L}{\partial\theta\partial\lambda'} \\ \frac{\partial^2 L}{\partial\alpha\partial\theta'} & \frac{\partial^2 L}{\partial\alpha\partial\alpha'} & \frac{\partial^2 L}{\partial\alpha\partial\lambda'} \\ \frac{\partial^2 L}{\partial\lambda\partial\theta'} & \frac{\partial^2 L}{\partial\lambda\partial\alpha'} & \frac{\partial^2 L}{\partial\lambda\partial\lambda'} \end{bmatrix} .$$

---

[1] The meaning of the assumption $E[\lambda_i] = 0$ is not clear. This lack of interpretation is not crucial, however, since (6.16) is introduced solely for notational purposes.

It is easily verified that the conformably partitioned inverse has the form

$$(6.19) \qquad \mathcal{R}_n^{-1} = \begin{bmatrix} N_n & * & * \\ * & M_n & * \\ * & * & * \end{bmatrix}$$

with M_n and N_n on the diagonal. Thus the inverse elements of a suitably bordered information matrix give lower bounds for the covariance matrices of any unbiased estimator of θ and α.

6.4. The Asymptotic Bounds

The asymptotic extension of Theorem 5 is straightforward. Defining

$$M = \lim \frac{1}{n} M_n \qquad \text{and} \qquad N = \lim \frac{1}{n} N_n$$

which are diagonal blocks of

$$\mathcal{R}^{-1} = \lim \begin{bmatrix} \frac{1}{n} R_n & 0 & -I \\ 0 & 0 & H' \\ -I & H & 0 \end{bmatrix}^{-1} = \begin{bmatrix} N & * & * \\ * & M & * \\ * & * & * \end{bmatrix},$$

one can state

THEOREM 6. In the presence of prior information expressed as (6.1), the matrices N and M are lower bounds for the asymptotic covariance matrices of any consistent estimators of θ and α. The efficiency gain

in estimating θ is given by $R^{-1} - N$, a positive semidefinite matrix of rank $m - r$. The bound is attained by the constrained maximum likelihood estimator.

Again the proof of the asymptotic theorem follows from applying the classical proofs of Theorem B. The minimum-chi-square estimators of θ and α are also asymptotically efficient. These are defined as the solution to the extremal problem

$$\min_{\theta} \quad (t - \theta)'\hat{R}(t - \theta)$$

$$\text{subject to} \quad \theta = h(\alpha)$$

where t is asymptotically normal with mean θ and covariance matrix R^{-1}.

6.5 Identification

It is of some interest to relax the assumption that the matrix H has full column rank. It is clear that the assumption has been used often in the above discussion since it is necessary for the inversion of the matrix $H'RH$. However, with some modifications, it is possible to generalize the results to handle the case where ρ is less than r. In the derivation of the lower bound for the covariance matrix of an estimator of θ a function $(c'y)^2$ was maximized over the set of all y such that $y = Hz$. That is, y was required to be in the column space of H. Suppose now that the columns of H are linearly dependent. Then the column space of H is spanned by a subset of the column vectors of H. Let H_1 be the matrix formed by such a subset. Then the constraint $y = Hz$ can be replaced by the constraint $y = H_1 z_1$. That is, the set of all y that can be written in the form $y = H_1 z_1$ for

some z_1 is identical to the set of all y that can be written as $y = Hz$ for some z. Hence equation (6.5) is valid if H is replaced by H_1. The a priori information increases the efficiency in estimating θ as long as ρ, the rank of H (and also the number of columns in H_1), is less than m, the number of rows of H.

The problem of estimating efficiently the parameter α is not so easily handled. Indeed, even when H has rank r there is a basic problem that we have not yet faced. The probability law for the sample is, according to our assumptions, uniquely determined by the parameter θ. If there is associated with the true parameter θ^o more than one vector α satisfying the equation

$$\theta^o = h(\alpha) , \tag{6.20}$$

it is not possible to speak of a "true" parameter α^o. Unless there is more a priori information concerning the parameter α, any solution of (6.20) is "true" in the sense that it implies the correct probability distribution of the observable sample. Hence estimation of α is possible only if h is a mapping such that θ^o has a unique image vector α^o.

This problem of estimating α when the likelihood function is in terms of θ is the essence of the famous "identification problem" in statistical inference. Although it would take us too far afield to discuss this problem at length here, it should be clear that the solution depends on the properties of the Jacobian matrix H. If H has rank r when evaluated at some vector α^o which satisfies (6.20), then α^o is at least locally unique.[1]

---

[1] Unfortunately, it need not be globally unique.

If, however, for some α^o satisfying (6.20) H has rank less than r, then there will exist an infinite number of solutions to (6.20).[1] Hence, if H nowhere has full column rank, it is clear that α is not identified. In such a case the constraint equations define new parameters α which cannot themselves be estimated. Nevertheless, as long as ρ is less than m the constraints sill impose restrictions on θ and are important in increasing the efficiency of estimating θ. The above discussion can be summarized in the following table which indicates the various possibilities.

	$\rho = r$	$\rho < r$
$\rho = m$	α locally identified no restrictions on θ	α not identified no restrictions on θ
$\rho < m$	α locally identified θ restricted	α not identified θ restricted

One further possibility that should be examined is that some elements of α may be identified but the others are not. Again the matrix H is the basis of the analysis. Suppose the vector α is partitioned into two parts, α_1 and α_2. Consider the equation[2]

$$(6.21) \qquad \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} \begin{bmatrix} d\alpha_1 \\ d\alpha_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

---

[1] This is true under the assumption made previously that the rank of H is constant in an open neighborhood of α^o. For a complete discussion of this point and the whole problem of unique identifiability, see Fisher [11].

[2] H is evaluated at some α^o which satisfies (6.20).

which is obtained by differentiating (6.20). The partitioning of H into two row blocks can be done in many different ways. Suppose, however, that there exists a partitioning such that H_{11} has full column rank and that H_{12} is a matrix of zeros. If this is the case, $d\alpha_1$ has the unique solution zero even if the full matrix H does not have full column rank. Since multiplying (6.21) by a nonsingular matrix is permissible, the following result can be stated. If H factors into AB where A is nonsingular and B is of the form

$$\begin{bmatrix} B_{11} & 0 \\ B_{21} & B_{22} \end{bmatrix} \tag{6.22}$$

with B_{11} having full column rank, then α_1 is locally identified. If α_1 is identified it may be estimated by an estimator whose covariance matrix is no less than the appropriate submatrix of $(H_1'RH_1)^{-1}$ where H_1 contains all the independent columns of H, including all those associated with α_1.

6.6. A Generalization

If we combine the two types of constraints considered in this and the previous section, a more general treatment is possible. Suppose, for example, that in addition to the set of constraints defining the new parameters α there are also constraints on α:

$$\begin{aligned} \theta_i &= h_i(\alpha) && (i = 1, \ldots, m) \\ g_i(\alpha) &= 0 && (i = 1, \ldots, k) . \end{aligned}$$

Again assuming that all the functions are continuously differentiable, we form the Jacobian matrix

$$\begin{bmatrix} -I & H \\ 0 & G \end{bmatrix} \tag{6.24}$$

where G and H are the $k \times r$ and $m \times r$ matrices

$$G = \left[\frac{\partial g_i}{\partial \alpha_j}\right] \qquad H = \left[\frac{\partial h_i}{\partial \alpha_j}\right] .$$

The parameter α will be locally identified if, when $\theta = \theta^0$,

$$\begin{bmatrix} H \\ G \end{bmatrix} d\alpha = 0 \tag{6.25}$$

has a unique solution zero. That is, α will be locally identified if $[H' \; G']$ has full row rank r. If in addition G has full row rank k, the bordered information matrix will possess an inverse:

$$\begin{bmatrix} R_n & 0 & -I & 0 \\ 0 & 0 & H' & G' \\ -I & H & 0 & 0 \\ 0 & G & 0 & 0 \end{bmatrix}^{-1} = \begin{bmatrix} \bar{N}_n & * & * & * \\ * & \bar{M}_n & * & * \\ * & * & * & * \\ * & * & * & * \end{bmatrix} \tag{6.26}$$

Using the methods of the preceding sections, one finds that the covariance matrix for any unbiased estimator of θ and α must satisfy the following inequalities:

$$\text{Var}\ \theta \geq N_n = HM_nH'$$

(6.27)

$$\text{Var}\ \alpha \geq M_n = Q - QG'(GQG')^{-1}GQ$$

where

(6.28) $$Q = (H'R_nH + G'G)^{-1} .$$

The asymptotic version of these inequalities is obvious. Again, the constrained maximum-likelihood estimator is asymptotically efficient and has a covariance matrix equal to the lower bound. The minimum-chi-square estimator, the solution of

$$\min_{\alpha} [t - h(\alpha)]'\hat{R}[t - h(\alpha)]$$

$$\text{subject to}\ \ g(\alpha) = 0 ,$$

is also asymptotically efficient if t is asymptotically normal and efficient without the constraints.

7. CASE OF INEQUALITY RESTRICTIONS

Perhaps the most natural way to express a priori information about the unknown parameter vector θ is by means of a set of inequalities. For example, one might know that θ_1 is greater than zero and that θ_2 lies between zero and one. In this section we explore the gain in estimation efficiency which can result from such constraints. Specifically, we shall assume that the statistician knows that the true parameter θ^o lies in A^*, an open subset of A. The assumption that A^* is open is made so that $f_n(x, \theta)$ will be differentiable at θ^o.

As in the previous three sections, it would seem that use of inequality constraints should reduce the Cramer-Rao lower bound for the variance of an unbiased estimator. In fact, however, this is not the case. The Cramer-Rao inequality was derived for an arbitrary open parameter space A. Since the lower bound does not depend on A, restricting A to A^* cannot lower the bound. The same argument also applies to the asymptotic Cramer-Rao inequality. Putting this more formally, we have

THEOREM 7. Despite the presence of a priori information which restricts θ to an open subset of A, the lower bound for the variance of an unbiased estimator of θ remains R_n^{-1}. Furthermore, the unconstrained maximum likelihood estimator remains asymptotically efficient with covariance matrix R^{-1}.

According to Theorem 7, inequality constraints are of no value in increasing efficiency as long as the requirement that estimators be unbiased is maintained. Furthermore, even if one drops the assumption of unbiasedness, it remains true that, asymptotically, inequality constraints do not help. For large n the probability that the maximum likelihood estimator violates the constraints is almost zero; hence there is nothing to gain by maximizing subject to constraint.

To interpret Theorem 7 to mean that inequality constraints are worthless would be a serious error. Rather, it points out that the classical theory, which is based on unbiasedness on the one hand and asymptotic approximations on the other, has important weaknesses. Nonstochastic prior

information according to this theory increases efficiency only if it reduces the dimensionality of the parameter space. Since inequality constraints do not reduce dimensionality, they cannot increase efficiency. But this result depends crucially on the way we have defined efficiency. A quite different answer would result under a more suitable definition which would remove the assumption of unbiasedness and simply be in terms of minimum mean squared error. Unfortunately, with such a definition, it can be shown that no efficient estimator exists independent of θ. However, it is possible to make some statements concerning the improvement of estimation precision under a priori information when the unbiasedness assumption is dropped.

Let us, for the rest of this section, use the following definition: An estimator s_n is better than an estimator t_n if

$$(7.1) \qquad E[(s_n - \theta)'B(s_n - \theta)] \leq E[(t_n - \theta)'B(t_n - \theta)]$$

for all positive definite (or semidefinite) matrices B and for all θ in the parameter space, with strict inequality occurring for some θ and B. Now suppose t_n is an unbiased estimator of θ which does not use the a priori information and which has a covariance matrix equal to the bound R_n^{-1}. Define the m-dimensional vector $z(t, B)$ to be that vector in $\bar{A}$, the closure of A^*, which is closest to t in the metric B. That is, z satisfies

$$(7.2) \qquad (t - z)'B(t - z) = \min_{x \in \bar{A}} (t - x)'B(t - x) .$$

Then consider the new estimator s_n defined as

$$s_n = \begin{cases} t_n & \text{if } t_n \epsilon \bar{A} \\ z & \text{if } t_n \notin \bar{A} \end{cases} \qquad (7.3)$$

The estimator s_n is equal to t_n if t_n happens to satisfy the constraints. If t_n does not satisfy the constraints, s_n equals that vector on the boundary of the constraint set which is "closest" to t_n. This new estimator will be a function of the sample X_n, the matrix B in the loss function (7.1), and the constraint set A^*. We shall prove in Appendix B the following result:

THEOREM 8. If the parameter θ is known to lie in an open convex set A^*, the estimator s_n is at least as good as the estimator t_n. Moreover, if $Pr[t_n \epsilon A^*]$ does not equal one, s_n is better than t_n.

The estimator s_n will of course be biased (otherwise it would violate Theorem 7) and will depend on the matrix of the quadratic loss function. Nevertheless it is an estimator that has smaller mean squared error than t_n, the estimator which attains the Cramer-Rao lower bound. Since s_n converges to t_n as n approaches infinity, there is no asymptotic version of Theorem 8 unless we again[1] accept the fiction that the a priori information increases with n. For small samples, however, considerable gain in efficiency may be possible.

---

[1] Cf. Section 4.3 above.

The assumption of convexity is restrictive, yet often met in practice. For example, the set of linear inequalities $C\theta > b$ or the positive definite quadratic form $(\theta - b)'C(\theta - b) < 1$ both give rise to open convex sets. Finding the estimator s_n will involve solving the quadratic programming problem in (7.2). This may in practice be quite difficult although some algorithms do exist. Finally, it should be noted that s_n is by no means an optimal estimator. All we have shown is that it is better than t_n. Without turning to Bayesian or minimax arguments, it is impossible to even define the problem of optimal estimation when the unbiasedness criterion is dropped.

8. SUMMARY

We have shown in the preceding sections how a priori information can be incorporated into the classical theory of estimation. In the first three cases both finite-sample and asymptotic results were possible. Unfortunately, the last case pointed out the difficulty of the classical approach. Finite-sample results depend crucially on the unbiasedness assumption, a criterion that is not easily defended. Asymptotic results are necessarily only approximately valid in application and the accuracy of the approximation is almost never known. Thus the results of this paper, like all those of classical statistics, must not be overstated.

A particularly interesting question arises in the case where the Cramer-Rao bound is attainable when no a priori information is present but the modified bound is not attainable with the a priori information. Suppose,

for example, one wishes to estimate the coefficient vector of the regression model (4.7). If no prior information is available the least-squares estimator is best. If the prior information is in the form of a nonlinear constraint equation, the bound P_n will not usually be attainable. But the maximum likelihood estimator will have a sampling distribution which can be approximated by a distribution which has a covariance matrix P. Since P is smaller than R^{-1}, one can conclude that the ML estimator $\hat{\beta}$ is better than the least-square estimator b as far as the approximation is valid. But the properties of b are known exactly whereas the properties of $\hat{\beta}$ are known only approximately. It is possible that b is better than $\hat{\beta}$ -- that using the a priori information on the basis of large-sample theory actually makes things worse. Thus one has the choice of using the estimator b which has known properties or using $\hat{\beta}$ which for large n is definitely better but for small n is perhaps worse.

The answer depends, of course, on how close the approximation is. But it also depends on our loss function. If our loss function is really an unbounded quadratic function (which is the basis presumably for minimum variance estimates), then any estimator whose distribution has thick enough tails will be rejected because of infinite variance. Yet the ML estimator under constraint may very well have infinite variance for every sample size n but, for large n, be approximated by a distribution with finite variance. What is needed to analyze these questions is a more careful theory of the appropriate (truncated, loss function. This issue is touched upon by Chernoff [6] but the problem remains unsettled. Until these issues are

better clarified in the statistical literature, all asymptotic results, including the ones given here, must be treated with caution. Nevertheless, it is probably useful to treat the asymptotic results as approximately valid until more evidence is available.

APPENDIX A

We shall derive here some of the properties of the covariance matrices which are discussed in this paper. Let R be an $m \times m$ positive definite matrix, G a $k \times m$ matrix having rank k, and H an $m \times r$ matrix having rank r. Consider the four matrices

$$A_1 = R^{-1} - P = R^{-1}G'(GR^{-1}G')^{-1}GR^{-1}$$

$$A_2 = P \qquad = R^{-1} - R^{-1}G'(GR^{-1}G')^{-1}GR^{-1}$$

$$A_3 = N \qquad = H(H'RH)^{-1}H'$$

$$A_4 = R^{-1} - N = R^{-1} - H(H'RH)^{-1}H'$$

We shall show that all four matrices are positive semidefinite with the rank of A_1 equal to k, the rank of A_2 equal to $m - k$, the rank of A_3 equal to r, and the rank of A_4 equal to $m - r$.

The derivation is based on the following facts:[1]

(1) Every positive definite matrix has a positive definite inverse.

(2) If A is positive definite, then there exists a nonsingular matrix D such that $A = D'D$.

(3) If D is nonsingular and B is positive semidefinite with rank ρ, then $D'BD$ is also positive semidefinite with rank ρ.

(4) If B is symmetric and idempotent (i.e., $B' = B = B^2$), then B is positive semidefinite with rank equal to the trace of B.

---

[1] See, for example, Graybill [12, pp. 1-17].

Since R^{-1} is positive definite it may be written as $D'D$. Then, defining the new matrices

$$X' = H'D^{-1}$$

$$Y' = GD' ,$$

we can write

$$A_1 = D'Y(Y'Y)^{-1}Y'D = D'B_1D$$

$$A_2 = D'D - D'Y(Y'Y)^{-1}Y'D = D'(I - B_1)D$$

$$A_3 = D'X(X'X)^{-1}X'D = D'B_2D$$

$$A_4 = D'D - D'X(X'X)^{-1}X'D = D'(I - B_2)D .$$

It is easy to verify that B_1, B_2, $I - B_1$, and $I - B_2$ are all idempotent and symmetric. Therefore, the A_i are positive semidefinite with the following ranks:

$$\rho(A_1) = tr[Y(Y'Y)^{-1}Y'] = tr[Y'Y)^{-1}Y'Y] = tr[I_k] = k$$

$$\rho(A_2) = tr[I_m] - tr[B_1] = m - k$$

$$\rho(A_3) = tr[B_2] = tr[(X'X)^{-1}X'X] = tr[I_r] = r$$

$$\rho(A_4) = tr[I_m] - tr[B_2] = m - r$$

where use has been made of the fact that $tr[AB] = tr[BA]$.

Another useful result concerning matrices is given by the following theorem:

THEOREM. If A is positive definite and B positive semidefinite, then $(A + B)^{-1}$ exists and is positive definite. Furthermore, $A^{-1} - (A + B)^{-1}$ is positive semidefinite having rank equal to the rank of B.

Proof. Let $C = A + B$. Then, for all vectors x,

$$x'Cx = x'Ax + x'Bx > 0 .$$

Thus C is positive definite. But all positive definite matrices possess positive definite inverses. By premultiplying by A and postmultiplying by C it is easily verified that

$$A^{-1} - C^{-1} = A^{-1}BC^{-1} .$$

Hence $A^{-1} - C^{-1}$ has rank equal to the rank of B.

Finally we must show that $A^{-1}BC^{-1}$ is positive semidefinite. Suppose first that B is nonsingular. Then

$$(A^{-1}BC^{-1})^{-1} = CB^{-1}A = (A + B)B^{-1}A = AB^{-1}A + A$$

is the sum of two positive definite matrices and is therefore positive definite. Thus its inverse $A^{-1}BC^{-1}$ must be positive definite.

If B is singular, consider the function

$$\lambda(\epsilon, x) = x'[A^{-1} - (A + B + \epsilon I)^{-1}]x$$

where ϵ is a scalar. Since $B + \epsilon I$ is positive definite for $\epsilon > 0$, we

know from the above that

$$\lambda(\epsilon, x) > 0$$

for all vectors x as long as $\epsilon > 0$. Furthermore, $\lambda(\epsilon, x)$ is continuous at $\epsilon = 0$. Hence $\lambda(0, x) \geq 0$ for all x. Thus, even if B is singular, $A^{-1} - C^{-1}$ is positive semidefinite.

APPENDIX B

We shall prove in this appendix Theorem 8 of Section 7. Consider first the following lemma:

Lemma 1. Let S be a closed convex set in n-dimensional Euclidean space. Let d be a distance function. Let t be a point exterior to S and let z be a point in S such that

$$d(z, t) \leq d(s, t) \tag{B.1}$$

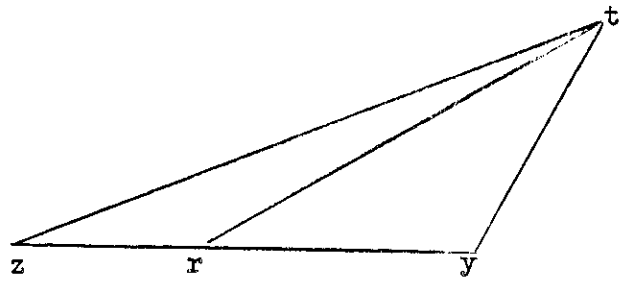
for all points s in S. Then,

$$d(z, x) < d(t, x) \tag{B.2}$$

for all x in S.

Proof: Suppose the lemma is false and that y is a point in S such that $d(z, y) \geq d(t, y)$. Then let r be that point on the line segment (z, y) such that $d(r, y) = d(t, y)$. Due to the convexity of S, r is in S and hence r is distinct from t. Construct the triangle which has vertices at r, t, and y. Since $d(t, y) = d(r, y)$ the angle try must be strictly less than 90°. (It may even be zero if r, t, and y are collinear). Hence the angle zrt must necessarily be greater than 90°. Thus the line segment (z, t) is the longest side of the obtuse triangle ztr. But this means that r is a point in S such that $d(r, t) < d(z, t)$. Since the assumption that $d(z, y) \geq d(t, y)$ leads to the violation of (B.1) we must

conclude that no such point y exists.



This lemma can be applied to the case considered in Theorem 8. Let S be $\bar{A}$ and

$$d(s, x) = (s - x)'B(s - x) \tag{B.3}$$

Then the point z defined by (7.2) satisfies (B.1). Recall from (7.3) that t_n and s_n are identical wherever t_n is in $\bar{A}$ and s_n equals z whenever t_n is exterior to $\bar{A}$. Since $\bar{A}$ is convex, it follows that

$$(t_n - \theta)'B(t_n - \theta) > (s_n - \theta)'B(s_n - \theta) \tag{B.4}$$

as long as t_n is exterior to $\bar{A}$. If this occurs with positive probability then

$$E(t_n - \theta)'B(t_n - \theta) > E(s_n - \theta)'B(s_n - \theta) \ . \tag{B.5}$$

Since the boundary of A is a set of measure zero, $Pr[t_n \in A^*]$ is the same as $Pr[t_n \in \bar{A}]$ and the theorem is proved.

REFERENCES

[1] Aitchison, J. and S. D. Silvey, "Maximum-likelihood estimation of parameters subject to restraints," Annals of Mathematical Statistics, Vol. 29 (1958), pp. 813-828.

[2] Aitchison, J. and S. D. Silvey, "Maximum-likelihood estimation procedures and associated tests of significance," Journal of the Royal Statistical Society, Series B, Vol. 22 (1960), pp. 154-171.

[3] Anderson, Theodore W., "Estimating linear restrictions on regression coefficients for multivariate normal distributions," Annals of Mathematical Statistics, Vol. 22, (1951) pp. 327-351.

[4] Basmann, R. L., "Remarks concerning the application of exact finite sample distribution functions of GCL estimators in econometric statistical inference," Journal of the American Statistical Association Vol. 58, (1963) pp. 943-976.

[5] Basmann, R. L., "On the application of the identifiability test statistic and its exact finite sample distribution function in predictive testing of explanatory economic models," mimeographed paper, (1965).

[6] Chernoff, Herman, "Large sample theory: parametric case," Annals of Mathematical Statistics, Vol. 27, (1956) pp. 1-22.

[7] Chiang, Chin Long, "On regular best asymptotically normal estimates," Annals of Mathematical Statistics, Vol. 27, (1956) pp. 336-351.

[8] Chipman, J. and M. M. Rao, "The treatment of linear restrictions in regression analysis," Econometrica, Vol. 32, (1964) pp. 198-209.

[9] Cramer, H., Mathematical Methods of Statistics, Princeton. Princeton University Press, 1946.

[10] Ferguson, T. S., "A method of generating best and asymptotically normal estimates with application to the estimation of bacterial densities," Annals of Mathematical Statistics, Vol. 29, (1958) pp. 1046-1062.

[11] Fisher, Franklin M., The Identification Problem in Econometrics, New York: McGraw-Hill, 1966.

[12] Graybill, Franklin A., An Introduction to Linear Statistical Models, Volume I, New York: McGraw-Hill, 1961.

[13] Goldberger, Arthur S., Econometric Theory, New York: John Wiley and Sons, 1964.

REFERENCES (continued)

[14] Goldberger, A. S., A. L. Nagar, and H. S. Odeh, "The covariance matrices of reduced-form coefficients and of forecasts for a structural econometric model," Econometrica, Vol. 29, (1961) pp. 556-573.

[15] Hammersley, J. N., "On estimating restricted parameters," Journal of the Royal Statistical Society, Series B, Vol. 12 (1950), pp.192-240.

[16] Kendall, M. G., and A. Stuart, The Advanced Theory of Statistics, Vol. 2, London, Griffin, 1961.

[17] Klein, L. R., Economic Fluctuations in the United States, 1921-1941, Cowles Commission Monograph 11, New York: Wiley, 1950.

[18] Klein, L. R., "The efficiency of estimation in econometric models," in Essays in Economics and Econometrics, Chapel Hill, University of North Carolina, 1960, pp. 216-232.

[19] Koopmans, T. C. and W. C. Hood, "The estimation of simultaneous linear economic relationships," Chapter 6 in Studies in Econometric Method, Cowles Commission Monograph 14, W. C. Hood and T. C. Koopmans, editors, New York: John Wiley and Sons, 1953, pp. 112-199.

[20] Koopmans, T. C., H. Rubin, and R. B. Leipnik, "Measuring the equation systems of dynamic economics," Chapter 2 in Statistical Inference in Dynamic Economic Models, Cowles Commission Monograph 10, T. C. Koopmans, editor, New York: John Wiley and Sons, 1950, pp. 53-237.

[21] LeCam, Lucien, "On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates," University of California Publications in Statistics, Vol. 1, No. 11, pp. 277-330.

[22] Malinvaud, E., Méthodes Statistique de l'Econométrie, Paris: Dunod, 1964.

[23] Rothenberg, T. J. and C. T. Leenders, "Efficient estimation of simultaneous equation systems," Econometrica, Vol. 32, (1964) pp. 57-76.

[24] Silvey, S. D., "The Lagrangian multiplier test," Annals of Mathematical Statistics, Vol. 30, (1959) pp. 389-407.

[25] Thrall, Robert M., and Leonard Tornheim, Vector Spaces and Matrices, New York: John Wiley and Sons, 1957.

[26] Theil, H., Economic Forecasts and Policy, Amsterdam: North-Holland, 1961.